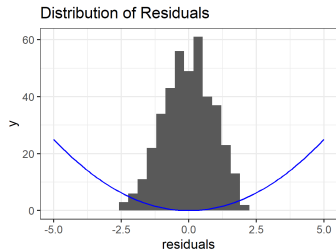


Introduction to Machine Learning

Maximum Likelihood Estimation vs. Empirical Risk Minimization



Learning goals

- Understand the connection between maximum likelihood and risk minimization
- Learn the correspondence between a Gaussian error distribution and the L2 loss

MAXIMUM LIKELIHOOD

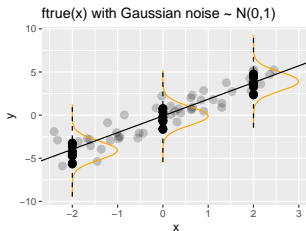
Let's consider regression from a maximum likelihood perspective.

Assume:

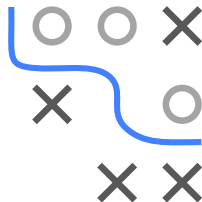
$$y \mid \mathbf{x} \sim p(y \mid \mathbf{x}, \theta)$$

Common case: true underlying relationship f_{true} with additive noise:

$$y = f_{\text{true}}(\mathbf{x}) + \epsilon$$



where f_{true} has params θ and ϵ a RV that follows some distribution \mathbb{P}_{ϵ} , with $\mathbb{E}[\epsilon] = 0$. Also, assume $\epsilon \perp\!\!\!\perp \mathbf{x}$.

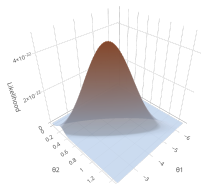


MAXIMUM LIKELIHOOD

From a statistics / maximum-likelihood perspective, we assume (or we pretend) we know the underlying distribution $p(y \mid \mathbf{x}, \theta)$.

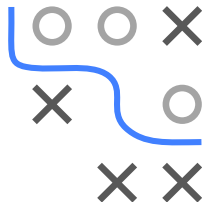
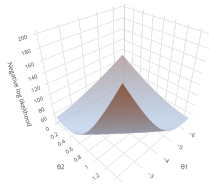
- Then, given i.i.d data $\mathcal{D} = ((\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(n)}, y^{(n)}))$ from \mathbb{P}_{xy} the maximum-likelihood principle is to maximize the **likelihood**

$$\mathcal{L}(\theta) = \prod_{i=1}^n p(y^{(i)} \mid \mathbf{x}^{(i)}, \theta)$$



or equivalently to minimize the **negative log-likelihood**

$$-\ell(\theta) = -\sum_{i=1}^n \log p(y^{(i)} \mid \mathbf{x}^{(i)}, \theta)$$



MAXIMUM LIKELIHOOD

From an ML perspective we assume our hypothesis space corresponds to the space of the (parameterized) f_{true} .

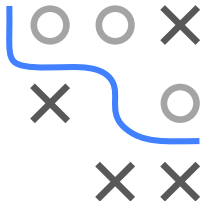
- Simply define neg. log-likelihood as **loss function**

$$L(y, f(\mathbf{x} \mid \theta)) := -\log p(y \mid \mathbf{x}, \theta)$$

- Then, maximum-likelihood = ERM

$$\mathcal{R}_{\text{emp}}(\theta) = \sum_{i=1}^n L(y^{(i)}, f(\mathbf{x}^{(i)} \mid \theta))$$

- NB: When we are only interested in the minimizer, we can ignore multiplicative or additive constants.
- We use \propto as “proportional up to multiplicative and additive constants”



GAUSSIAN ERRORS - L2-LOSS

Assume $y = f_{\text{true}}(\mathbf{x}) + \epsilon$ with additive Gaussian errors, i.e. $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$. Then

$$y \mid \mathbf{x} \sim N(f_{\text{true}}(\mathbf{x}), \sigma^2)$$

The likelihood is then

$$\begin{aligned}\mathcal{L}(\theta) &= \prod_{i=1}^n p\left(y^{(i)} \mid f(\mathbf{x}^{(i)} \mid \theta), \sigma^2\right) \\ &\propto \prod_{i=1}^n \exp\left(-\frac{1}{2\sigma^2} \left(y^{(i)} - f(\mathbf{x}^{(i)} \mid \theta)\right)^2\right)\end{aligned}$$



GAUSSIAN ERRORS - L2-LOSS

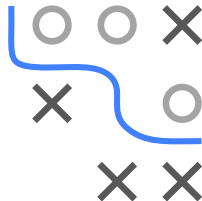
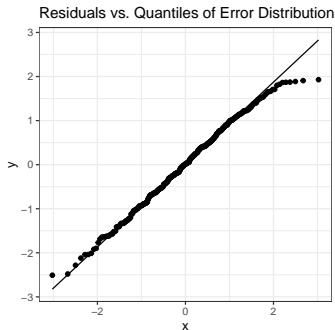
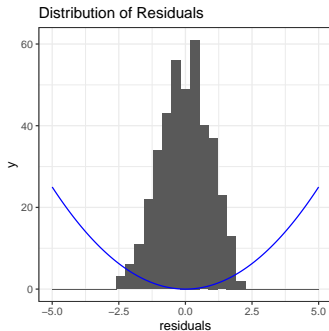
Easy to see: minimizing neg. negative log-likelihood with Gaussian errors is the same as ERM with L_2 -loss:

$$\begin{aligned}-\ell(\boldsymbol{\theta}) &= -\log(\mathcal{L}(\boldsymbol{\theta})) \\ &= -\log\left(\prod_{i=1}^n \exp\left(-\frac{1}{2\sigma^2} \left(y^{(i)} - f(\mathbf{x}^{(i)} | \boldsymbol{\theta})\right)^2\right)\right) \\ &\propto \sum_{i=1}^n \left(y^{(i)} - f(\mathbf{x}^{(i)} | \boldsymbol{\theta})\right)^2\end{aligned}$$



GAUSSIAN ERRORS - L2-LOSS

- We simulate data $y \mid \mathbf{x} \sim \mathcal{N}(f_{\text{true}}(\mathbf{x}), 1)$ with $f_{\text{true}} = 0.2 \cdot \mathbf{x}$
- Let's plot empirical errors as histogram, after fitting our model with L_2 -loss
- Q-Q-plot compares empirical residuals vs. theoretical quantiles of Gaussian



DISTRIBUTIONS AND LOSSES

- For every error distribution \mathbb{P}_ϵ we can derive an equivalent loss function, which leads to the same point estimator for the parameter vector θ as maximum-likelihood. Formally,
 - $\hat{\theta} \in \arg \max_{\theta} \mathcal{L}(\theta) \implies \hat{\theta} \in \arg \min_{\theta} -\log(\mathcal{L}(\theta))$
- **But:** The other way around does not always work: We cannot derive a corresponding pdf or error distribution for every loss function – the Hinge loss is one prominent example, for which some probabilistic interpretation is still possible however, see

► Sollich, 1999 .



DISTRIBUTIONS AND LOSSES

When does the reverse direction hold?

- If we can write the loss as $L(y, f(\mathbf{x})) = L(y - f(\mathbf{x})) = L(r)$ for $r \in \mathbb{R}$, then minimizing $L(y - f(\mathbf{x}))$ is equivalent to maximizing a conditional log-likelihood $\log(p(y - f(\mathbf{x}|\theta)))$ if
 - $\log(p(r))$ is affine trafo of L (undoing the \propto):

$$\log(p(r)) = a - bL(r), \quad a \in \mathbb{R}, b > 0$$

- p is a pdf (non-negative and integrates to one)

Thus, a loss L corresponds to MLE under *some* distribution if there exist $a \in \mathbb{R}$, $b > 0$ such that

$$\int_{\mathbb{R}} \exp(a - bL(r)) dr = 1$$

